Slovene Twitter Analytics

Nikola Ljubešić*†, Darja Fišer†‡

* Dept. of Knowledge Technologies, Jožef Stefan Institute
† Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb
‡ Dept. of Translation, Faculty of Arts, University of Ljubljana

CMC conference 2016, 27th Sep 2016
Task

Twitter

- Large volumes of metadata-rich CMC data
- Data easily obtainable
- Reasonable terms of service
## Twitter
- Large volumes of metadata-rich CMC data
- Data easily obtainable
- Reasonable terms of service

## Data analytics
- Explorative analysis of all / many variables available
- Very popular recently ("data science" buzzword) as
  - Most datasets were not produced by researchers, so variables were not controlled
  - Many datasets are crossreferenced – immense number of variables
Rios and Lin (2013) analyse annual tweeting dynamics around the world – identifying interesting cultural differences.

Scheffler and Kyba (2016) investigate morning routine of German Twitter users – social norms of working life.
Related work

- Rios and Lin (2013) analyse annual tweeting dynamics around the world – identifying interesting cultural differences
- Scheffler and Kyba (2016) investigate morning routine of German Twitter users – social norms of working life
- Bamman et al. (2012) analyse dependence of gender and
  - language standardness (women more predominant)
  - communication style (men informative, women involved)
  - vocabulary (women more distinct features)
- Arakawa et al. (2014) discriminate between organisations and private users, most others discriminate between more user types
JANES corpus

- Tweets, fora, blogs, news comments, Wikipedia talk pages
- 7.5 million tweets, 107 million tokens posted by ~9,000 users between June 2013 and January 2016
Dataset

JANES corpus
- Tweets, fora, blogs, news comments, Wikipedia talk pages
- 7.5 million tweets, 107 million tokens posted by ~9,000 users between June 2013 and January 2016

Data collection
- Collected with TweetCat (Ljubešić et al., 2014) – based on Search Twitter API (https://github.com/nljubes/tweetcat)
- Part of an emerging toolkit
  - TweetGeo for collecting geo-encoded tweets published in an area
  - TweetPub for preparing your data for publishing
Variables of interest

Metadata-based
- timestamp of publishing
- was the tweet retweeted
- was the tweet favourited
Variables of interest

<table>
<thead>
<tr>
<th>Metadata-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>- timestamp of publishing</td>
</tr>
<tr>
<td>- was the tweet retweeted</td>
</tr>
<tr>
<td>- was the tweet favourited</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>- text standardness (Ljubešić et al., 2015)</td>
</tr>
<tr>
<td>- sentiment (Fišer et al., 2016)</td>
</tr>
</tbody>
</table>
Variables of interest

**Metadata-based**
- timestamp of publishing
- was the tweet retweeted
- was the tweet favourited

**Text-based**
- text standardness (Ljubešić et al., 2015)
- sentiment (Fišer et al., 2016)

**User-level**
- gender of the user – male or female
- account type – private or corporate
Overview

Posting dynamics

- Daily and weekly
- Dependence on the gender and account type variables
Overview

Posting dynamics

- Daily and weekly
- Dependence on the gender and account type variables

Retweets and favorites

- Dependence on the gender and account type variables
Overview

Posting dynamics
- Daily and weekly
- Dependence on the gender and account type variables

Retweets and favorites
- Dependence on the gender and account type variables

Language standardness
- Dependence on the gender and account type variables
- Daily dynamics
## Overview

### Posting dynamics
- Daily and weekly
- Dependence on the gender and account type variables

### Retweets and favorites
- Dependence on the gender and account type variables

### Language standardness
- Dependence on the gender and account type variables
- Daily dynamics

### Sentiment
- Dependence on the gender and account type variables
- Weekly dynamics
- Dependence on the standardness variable
Weekly posting dynamics

- **Dataset**
- **Results**
- **Conclusion**
Daily posting dynamics

- **Private**
- **Corporate**

- **Male**
- **Female**

0.00 0.02 0.04 0.06 0.08
## Retweets and favorites

<table>
<thead>
<tr>
<th></th>
<th>retweeted</th>
<th>favorited</th>
</tr>
</thead>
<tbody>
<tr>
<td>private</td>
<td>8.5%</td>
<td>30.2%</td>
</tr>
<tr>
<td>corporate</td>
<td>16.3%</td>
<td>18.0%</td>
</tr>
<tr>
<td>male</td>
<td>9.4%</td>
<td>29.2%</td>
</tr>
<tr>
<td>female</td>
<td>6.8%</td>
<td>32.9%</td>
</tr>
</tbody>
</table>
Retweets and favorites

<table>
<thead>
<tr>
<th></th>
<th>retweeted</th>
<th>favored</th>
</tr>
</thead>
<tbody>
<tr>
<td>private</td>
<td>8.5%</td>
<td>30.2%</td>
</tr>
<tr>
<td>corporate</td>
<td>16.3%</td>
<td>18.0%</td>
</tr>
<tr>
<td>male</td>
<td>9.4%</td>
<td>29.2%</td>
</tr>
<tr>
<td>female</td>
<td>6.8%</td>
<td>32.9%</td>
</tr>
</tbody>
</table>

source vs. retweeted $X^2(1, N = 7503200) = 74308, p < .001$
Retweets and favorites

<table>
<thead>
<tr>
<th></th>
<th>retweeted</th>
<th>favorited</th>
</tr>
</thead>
<tbody>
<tr>
<td>private</td>
<td>8.5%</td>
<td>30.2%</td>
</tr>
<tr>
<td>corporate</td>
<td>16.3%</td>
<td>18.0%</td>
</tr>
<tr>
<td>male</td>
<td>9.4%</td>
<td>29.2%</td>
</tr>
<tr>
<td>female</td>
<td>6.8%</td>
<td>32.9%</td>
</tr>
</tbody>
</table>

- source vs. retweeted $X^2(1, N = 7503200) = 74308, p < .001$
- source vs. favorited $X^2(1, N = 7503200) = 80215, p < .001$
## Retweets and favorites

<table>
<thead>
<tr>
<th></th>
<th>Retweeted</th>
<th>Favorited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private</td>
<td>8.5%</td>
<td>30.2%</td>
</tr>
<tr>
<td>Corporate</td>
<td>16.3%</td>
<td>18.0%</td>
</tr>
<tr>
<td>Male</td>
<td>9.4%</td>
<td>29.2%</td>
</tr>
<tr>
<td>Female</td>
<td>6.8%</td>
<td>32.9%</td>
</tr>
</tbody>
</table>

- Source vs. retweeted $X^2(1, N = 7503200) = 74308, p < .001$
- Source vs. favorited $X^2(1, N = 7503200) = 80215, p < .001$
- Gender vs. retweeted $X^2(1, N = 7503200) = 11714, p < .001$
## Retweets and favorites

<table>
<thead>
<tr>
<th></th>
<th>retweeted</th>
<th>favored</th>
</tr>
</thead>
<tbody>
<tr>
<td>private</td>
<td>8.5%</td>
<td>30.2%</td>
</tr>
<tr>
<td>corporate</td>
<td>16.3%</td>
<td>18.0%</td>
</tr>
<tr>
<td>male</td>
<td>9.4%</td>
<td>29.2%</td>
</tr>
<tr>
<td>female</td>
<td>6.8%</td>
<td>32.9%</td>
</tr>
</tbody>
</table>

- source vs. retweeted $X^2(1, N = 7503200) = 74308, p < .001$
- source vs. favorited $X^2(1, N = 7503200) = 80215, p < .001$
- gender vs. retweeted $X^2(1, N = 7503200) = 11714, p < .001$
- gender vs. favorited $X^2(1, N = 7503200) = 8913.4, p < .001$
Standardness vs. source and gender

24% of female and 19.6% of male tweets very non-standard
**Standardness vs. source and gender**

- **gender vs. standardness**
  \[ X^2(1, N = 7503200) = 9740.9, p < .001 \]
- 24% of female and 19.6% of male tweets very non-standard
Standardness and daily dynamics
Sentiment vs. source and gender

![Bar chart showing sentiment vs. source and gender](chart.png)
sentiment vs. gender $X^2(1, N = 7503200) = 6179.8, p < .001$
Sentiment vs. text standardness

The diagram shows the distribution of sentiment across different text standardness levels (L1, L2, L3). The x-axis represents the sentiment categories: negative, neutral, and positive. The y-axis represents the frequency or proportion of each sentiment type.

- **L1**: A moderate distribution across negative, neutral, and positive sentiments.
- **L2**: A higher concentration of positive sentiment compared to the other two categories.
- **L3**: The highest proportion of positive sentiment, with a notable difference from the other categories.

The chart illustrates that as the text standardness increases (L1 to L3), the proportion of positive sentiment also increases.
Weekly dynamics of sentiment

- Monday: Negative sentiment is highest, followed by neutral and then positive.
- Tuesday: Similar trend as Monday, with negative sentiment being the highest.
- Wednesday: Similar to Monday and Tuesday, with negative sentiment leading.
- Thursday: Consistent trend, negative sentiment is the highest.
- Friday: Similar pattern, negative sentiment is the highest.
- Saturday: Negative sentiment is the highest again.
- Sunday: Negative sentiment is slightly lower than other days, with neutral sentiment being the highest.

The chart shows that throughout the week, negative sentiment is the most common, followed by neutral and then positive sentiment.
Conclusion

- Explorative analysis of a series of extralinguistic and linguistic variables
Conclusion

- Explorative analysis of a series of extralinguistic and linguistic variables
- Big differences between tweeting behaviour, content and treatment of corporate and private tweets – in line with related work
Conclusion

- Explorative analysis of a series of extralinguistic and linguistic variables
- Big differences between tweeting behaviour, content and treatment of corporate and private tweets – in line with related work
- Non-standard language used mostly in the “early” hours
- Male users tweet more during weekdays and mornings
- Female users tweet more positive, the prevailing sentiment during weekends
- Male users use more standard language – contrary to findings of Bamman et al. (2012)
Conclusion

- Explorative analysis of a series of extralinguistic and linguistic variables
- Big differences between tweeting behaviour, content and treatment of corporate and private tweets – in line with related work
- Non-standard language used mostly in the “early” hours
- Male users tweet more during weekdays and mornings
- Female users tweet more positive, the prevailing sentiment during weekends
- Male users use more standard language – contrary to findings of Bamman et al. (2012)
- Future work
  - More variables, more sources, more languages
Explorative analysis of a series of extralinguistic and linguistic variables

Big differences between tweeting behaviour, content and treatment of corporate and private tweets – in line with related work

Non-standard language used mostly in the “early” hours

Male users tweet more during weekdays and mornings

Female users tweet more positive, the prevailing sentiment during weekends

Male users use more standard language – contrary to findings of Bamman et al. (2012)

Future work
  - More variables, more sources, more languages
  - Hypotheses, in-depth analysis and comparison to related work
Slovene Twitter Analytics

Nikola Ljubešić*, Darja Fišer†*

* Dept. of Knowledge Technologies, Jožef Stefan Institute
† Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb
‡ Dept. of Translation, Faculty of Arts, University of Ljubljana

CMC conference 2016, 27th Sep 2016